**STA 502:** Statistical Programming

**Final SAS Project Report**

**Author:** Kelvin Njuki

**Due Date:** 12/13/2019

**Association Between Health-Related Risk Behaviors and Chronic Diseases in United States**

**INTRODUCTION**

Based on literature and past findings, health-related risk behaviors such as smoking, excessive drinking of alcohol among others, were found to have some association with chronic diseases such as cancer, heart attack among others. This project focused on ascertaining these findings in United States. 2018 data collected by Center for Disease Control (CDC) using Behavioral Risk Factor Surveillance System (BRFSS) was used to conduct analysis. Data was about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The project mostly focused on cancer, asthma, diabetes and arthritis as chronic diseases and smoking, binge drinking and use of smokeless tobacco products as health-related risk behaviors.

**OBJECTIVES**

Major objective was to find out whether there exist any interesting association between health-related risk behaviors and chronic diseases in United States. Other objectives were:

1. To assess the association between chronic diseases and age categories in United States.
2. To find out whether there exist an association between health-related risk behaviors and chronic diseases in each state.
3. To find out whether there exist any association between health-related risk behaviors and chronic diseases in each state to a given age category.

The analysis was conducted using SAS Statistical package as outlined below.

**DATA CLEANING STEPS**

Data was obtained from CDC website (<https://www.cdc.gov/brfss/annual_data/2018/files/LLCP2018XPT.zip>) as XPT transport file. Conversion was done to SAS dataset (.sas7bdat) using a code provide from (<https://www.cdc.gov/brfss/annual_data/2018/files/TRANSPRT.SAS>) with slight modification as guided. This SAS code was submitted for grading ([FinalProjectDataConversionCode.sas](file:///E:\2.Miami\ClassWork\Fall2019\STA%20502%20-%20Statistical%20Programming\STA%20502\Final%20SAS%20Project\Finals\FinalProjectDataConversionCode.sas)). SAS library was created for storage of the SAS dataset. After learning that data had so many missing values, a quick exploration was done using PROC MEANS procedure to find out magnitude of missingness in every variable. By good luck, variables of interest had fewer missing values. State was one of variable of interest, but data contained only state fips codes. Therefore, a different dataset with state fips codes with their respective state names was downloaded from (<http://staff.washington.edu/glynn/StateFIPSicsprAB.xls>) and saved as a csv file. The data was imported in SAS and merged with the large dataset from BRFSS by state fips codes. This was as a quicker way of supplying state names that corresponds to respective state fips codes.

With over 270 variables, the project required only 15 variables to achieve the set objectives. Therefore, a subset of data was generated using PROC SQL. Subset data contained only 15 variables of interest.

**GENERATING PLOTS AND SUMMARY TABLES**

Having concluded data cleaning steps, a few SAS MACRO were built to allow flexibility and efficiency in computational evaluations of the dataset. The MACROs were broken down into simple MACROs handling each of the above stated objectives. However, an overall MACRO was built in case the user may be interested in generating report for the four MACROs at once saving it in a word document. Below are list of the MACROs built and their purpose;

1. A SAS MACRO to output association between chronic diseases and health related risk behaviors in United States. This MACRO allows user to input chronic disease and health related risk behavior of interest. Cancer and binge drinking were used when calling MACRO for demonstration purpose.
2. A SAS MACRO to output association between age category of respondents and chronic disease. Cancer was used in the code for demonstration purpose. Age categories are in groups of 5 years.
3. A SAS MACRO to output association between chronic disease and health related risk behaviors in each state. The MACRO can subset data according to the state name supplied. The subset data is then used to generate a contingency table, mosaic plots and statistics table for categorical variables supplied. This can help each state to understand whether there exist any association between chronic diseases and health related risk behaviors.
4. SAS MACRO to investigate the association between chronic disease and heath related risk behavior in each state in each age category. This MACRO will help state users to investigate each of the age categories.
5. An overall SAS MACRO to combine the above four MACROs, generate a report and save it in a word document.
6. SAS MACRO for non-parametric bootstrap of brfss\_clean data, generate p-values and calculate the proportion of p-values that leads to rejection of null hypothesis. In this MACRO, user can be able to set seed of interest, number of replications interested with and has freedom to choose output data name of choice. The MACRO is aiming at confirming results from MACRO (i).
7. SAS MACRO to conduct permutation test where chi-square failed due to less than 5 expected counts.

**RESULTS**

The project focused on testing below hypotheses

**H0:** No association exists between health-related risk behaviors and chronic diseases

**H1:** There exists association between health-related risk behaviors and chronic diseases

1. ***MACRO (i) Results: Association between chronic diseases and health related risk behaviors in United States***



**Figure 1: Distribution of binge drinking and any other type of cancer apart from skin cancer**

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 6 | 2431.6954 | <.0001 |
| **Likelihood Ratio Chi-Square** | 6 | 2492.8746 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 526.5789 | <.0001 |
| **Phi Coefficient** |  | 0.0746 |  |
| **Contingency Coefficient** |  | 0.0744 |  |
| **Cramer's V** |  | 0.0527 |  |

**Table 1: Statistics for Table of Binge drinking by any other type of disease apart from skin cancer**

The chi-square p-value is **0.0001** which is less than **0.05**. Therefore, we reject null hypothesis and conclude that there exists an association between binge drinking and being diagnosed with any other type of cancer apart from skin cancer. These results were ascertained by conducting a non-parametric bootstrapping ***(MACRO (vi))*** which gave below results;

| **Analysis Variable : reject** |
| --- |
| **Mean** |
| 1.0000000 |

**Table 2: Proportion of simulations where null hypothesis was rejected after non-parametric bootstrapping**

Only 100 replications were ran due to storage space in SAS, user can supply number of reps of interest in the macro variable. Bootstrapping reports that out of 100 simulations, all resulted to rejection of null hypothesis.

1. ***MACRO (ii) Results: Association between chronic diseases and age categories in United States***



**Figure 2: Distribution of any other type of cancer apart from skin cancer by age categories**

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 39 | 25812.2374 | <.0001 |
| **Likelihood Ratio Chi-Square** | 39 | 25145.1018 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 7986.6681 | <.0001 |
| **Phi Coefficient** |  | 0.2429 |  |
| **Contingency Coefficient** |  | 0.2361 |  |
| **Cramer's V** |  | 0.1403 |  |

**Table 3: Statistics for Table of any other type of cancer apart from skin cancer and age categories**

P-value **(0.0001)** is less than **0.05** leading to rejection of null hypothesis. Data provided enough evidence to conclude that there exist an association between age categories and being diagnosed with any other type of cancer apart from skin cancer.

1. ***MACRO (iii) Results: Association between chronic diseases and health related risk behaviors in New York State***



**Figure 3: Distribution of binge drinking and any other type of cancer apart from skin cancer in Alabama State**

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 6 | 34.7776 | <.0001 |
| **Likelihood Ratio Chi-Square** | 6 | 40.3057 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 0.1512 | 0.6974 |
| **Phi Coefficient** |  | 0.0727 |  |
| **Contingency Coefficient** |  | 0.0725 |  |
| **Cramer's V** |  | 0.0514 |  |
| **WARNING: 42% of the cells have expected counts less than 5. Chi-Square may not be a valid test.** | | | |

**Table 4: Statistics for Table of binge drinking and any other type of cancer apart from skin cancer in Alabama State**

MACRO was tested using Alabama state data. P-value was less than 0.05 which suggest we reject null hypothesis. However, the warning message tells us that chi-square test was not the best due to less than 5 expected counts. Further step was taken, permutation test was done ***(MACRO (vii))*** and yielded below results;

| **Analysis Variable : rejection** |
| --- |
| **Mean** |
| 0.0570000 |

**Table 5: Proportion of simulation where null hypothesis was rejected in after conducting permutation test**

This implies that out of 1000 simulations, we rejected null hypothesis 57 time. This is a very small proportion; therefore our decision is failing to reject null hypothesis. We conclude that in Alabama State, there exist no association between binge drinking and being diagnosed with any other type of cancer apart from skin cancer. Similar warning may be experienced in ***MACRO (iv),*** a slight modification of ***MACRO (vii)*** will help confirm the results as we have done above.

One thing you may have noticed, titles in the reports generated do not look clear to the user. Variables names appearing on the title may be difficult to understand. I tried my level best to find out how I can use variable labels instead but it didn’t work.